High-speed networking for scientific applications

PhD students: Zhenzhen Yan, Jie Li, Zhengyang Liu, Tian Jin, Zhe Song

Collaborators: Chris Tracy, ESnet, Steve Emmerson, UCAR, John Dennis, NCAR, Robert D. Rusell, UNH, David Starobinski, BU

Malathi Veeraraghavan
University of Virginia
mvee@virginia.edu
Jan. 15, 2013

Web site: http://www.ece.virginia.edu/mv/html-files/research.html

Thanks to the US DOE ASCR for grants DE-SC002350 and DE-SC0007341 & NSF for grants, OCI-1127340 , OCI-1038058, and CNS-1116081

Outline

- Background
  - National labs & Res-and-Edu Nets (RENs)
  - What's special about scientific applications?
- UVA research projects
  - Large dataset movement
  - Hybrid network traffic engineering system
  - Virtual circuit multicast transport protocol
  - Scheduled Circuit Routing Protocol
  - Datacenter networking

Scientific community

- Department of Energy (DOE) funds fundamental research in basic sciences
  - High energy physics
  - Basic energy sciences
  - Biological and Environmental Research
  - Fusion Energy Sciences
  - Nuclear Physics
  - Advanced Scientific Computing Research (ASCR)
- National Science Foundation Office of Cyber Infrastructure (NSF OCI)
  - University Corporation for Atmospheric Research (UCAR)
  - National Center for Atmospheric Research (NCAR)
  - Teragrid and now XSEDE

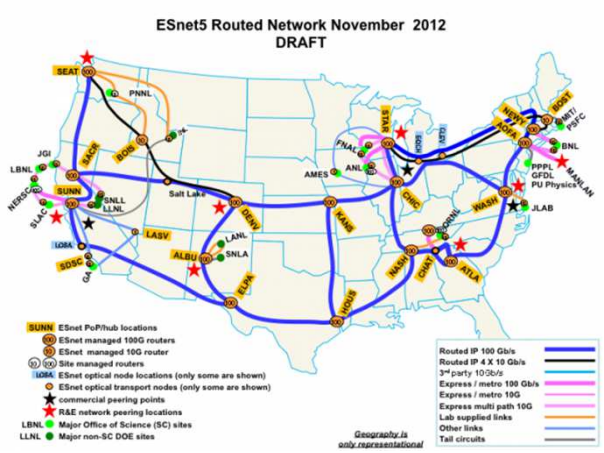
3

Labs, universities, computing, & Research & Education Networks (REN)

- DOE national labs
  - Oak Ridge National Lab (ORNL)
  - Argonne National Lab (ANL)
  - Lawrence Berkeley National Laboratory (LBNL)
  - National Energy Research Scientific Computing Center (NERSC)
  - and other labs …
- Universities – Physicists, Biologists, Climatologists, etc.
- Supercomputing facilities
  - Argonne Leadership Computing Facility (ALCF), OLCF (ONRL), NERSC
  - TACC (Texas), NWSC (NCAR Wyoming Supercomputing Center), etc.
- ESnet: 100 Gb/s backbone REN for DOE labs
- Internet2: Backone REN for universities/labs

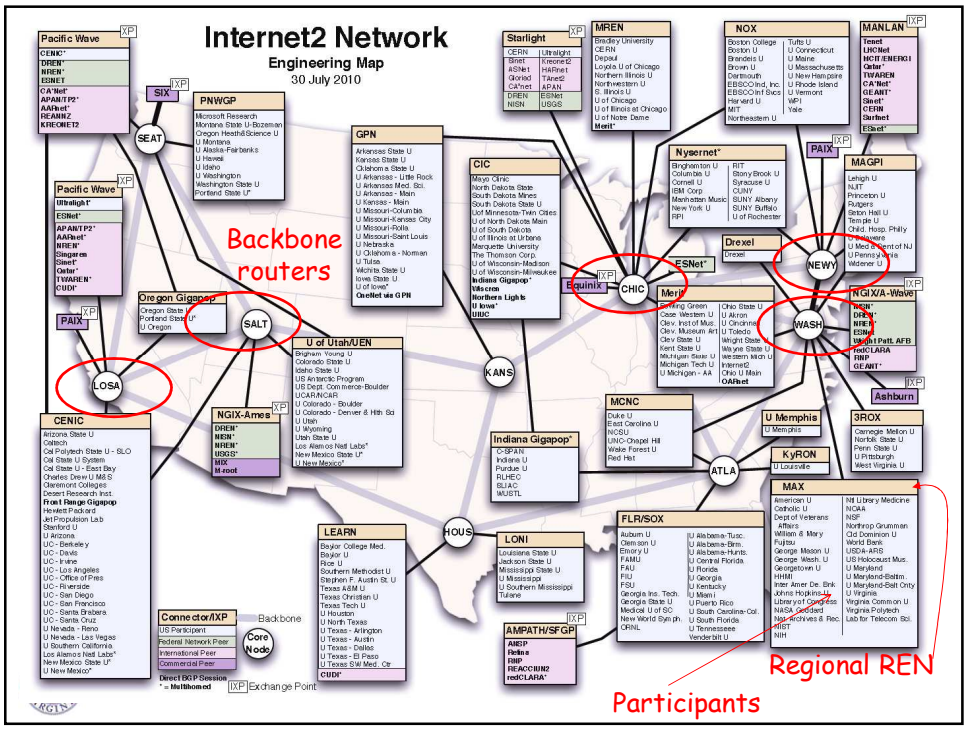
4

Energy Sciences Network (ESnet)



Network page in http://es.net

5



Links of interest

- http://www.internet2.edu/network
- http://www.hawaii.edu/internet2/
- Global NOC (Indiana University)
  - http://globalnoc.iu.edu (supported networks)
    Try Internet2 – see live atlas
- Routerproxy
  - http://routerproxy.grnoc.iu.edu/internet2
  - allows user to execute JunOS commands
  - obtain routing tables, traceroute
  - useful teaching tool

7

Other RENs

- GEANT2 (European)
- JGN-X (Japan)
- Canarie (Canada)
- RNP (Brazil)
- China Education and Research Network (CERNET2)

8

Outline

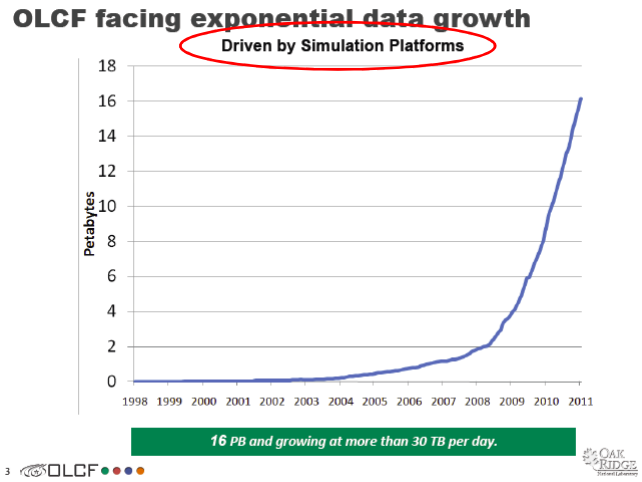
- Background
  - National labs & Res-and-Edu Nets (RENs)
  - ➢What's special about scientific applications?
- UVA research projects
  - Large dataset movement
  - Hybrid network traffic engineering system
  - Virtual circuit multicast transport protocol
  - Scheduled Circuit Routing Protocol
  - Datacenter networking

Scientific applications

- Instruments (create large datasets):
  - Particle accelerators (Large Hadron Collider), Telescopes, Satellites, Radar sites, Argonne Photon Source, Very-long baseline interferometry (VLBI), Very Large Array (VLA), Laser interferometer Gractivational wave Observatory (LIGO), ITER
- Experiments: ATLAS, CMS on LHC (15 PB annually)
- Sequencing: e.g., Genomics (large volumes)
  - 1000-Genome 200 TB dataset on Amazon EC2
- Model-based reanalyses: hybrid model-observational data sets
- Scientific Discovery through Advanced Computing (SCIDAC)
  - Volumes of simulation data
  - Petaflops → exaflops ⇒ larger storage ⇒ drive to Tb/s networks

J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling, "Climate data challenges in the 21st century," Science, vol. 331, no. 6018, pp. 700–702, 2011.

Growth in required storage at Oak Ridge Leadership Computing Facility (OLCF)



Galen Shipman, ORNL, DOE Terabits Workshop, Feb. 2011

11

An example supercomputing center

- http://www.nersc.gov/systems
- Hopper Cray XE6
- Edison Cray XC30
- Carver IBM iDataPlex (InfiniBand cluster)
- HPSS data archive (storage)
- Data transfer nodes
- Dirac: GPU computing

12

Communication needs

- Large dataset movement
  - Higher the rate, the better. Relentless!
  - Throughput variance has been a concern
- Remote visualization
  - Big data needs analytics & viz
  - OpenGL commands – not high rate, but low latency
- Remote instrument control
  - Beamline control of Photon source
  - Haptics, remote surgery, etc.
  - Low jitter
- Remote computational steering
  - Reduce overhead of running simulations, downloading data, changing parameters, rerunning simulations, etc.

Outline

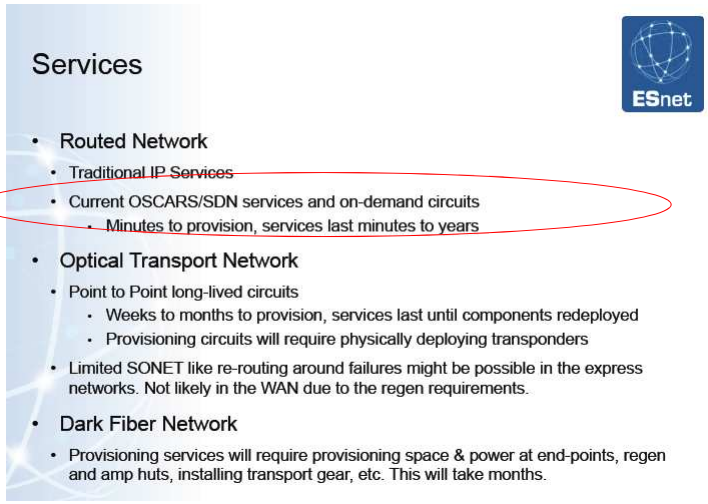
- Background
  - National labs & Res-and-Edu Nets (RENs)
  - What's special about scientific applications?
- ➢ UVA research projects
  - Large dataset movement
  - Hybrid network traffic engineering system
  - Virtual circuit multicast transport protocol
  - Scheduled Circuit Routing Protocol
  - Datacenter networking

Common themes

- Dynamic virtual circuits
- Data analysis
  - science before engineering
  - Theodore von Karmann paraphrased:
    - Scientists discover that which is
    - Engineers create that which never was

15

ESnet5



Joe Metzger, ESnet, ESCC, Jan. 2012

16

Lots of network data available

- NetFlow data
  - REN providers give researchers anonymized NetFlow data (1-in-1000 sampling: header data)
- SNMP data (bits/sec)
- Application usage statistics
  - GridFTP usage logs (> 2000 servers)
  - UCAR climate data distribution logs
- PerfSONAR (active measurements)
  - One-Way Ping (OWAMP) delay
  - BWCTL (Throughput)

17

UVA research projects

- Large dataset movement: GridFTP
  - ESnet wants to separate these flows from other flows, and use traffic-engineered paths
  - VC setup offers opportunity for path selection (traffic engineering), and for isolating flows to their own queues
- Hybrid network traffic engineering system
- Virtual circuit multicast transport protocol

18

GridFTP usage log analysis

- Purpose: characterize wide area data movement
- Obtained logs from
  - NERSC-ORNL (Sept. 2010)
  - NERSC-ANL (Mar. 4 - Apr. 22, 2012)
  - NCAR-NICS (2009-2011)
  - SLAC-BNL (Feb. 10 - Apr. 26, 2012)
- Analysis performed
  - Session Analysis
  - Throughput Variance Analysis

Z. Liu, M. Veeraraghavan, Z. Yan, C. Tracy, J. Tie, I. Foster, J. Dennis, J. Hick, Y. Li, and W. Yang, "On using virtual circuits for GridFTP transfers," SC 2012.

19

Format of a transfer log

```
1 DATE=20121014120031.536839 HOST=dtn02.nersc.gov PROG=globus-gridftp-server
   NL.EVNT=FTP_INFO START=20121014120003.275074 USER=zl4ef FILE=/dev/zero BU
  FFER=1048576 BLOCK=262144 NBYTES=21494235136 VOLUME=/ STREAMS=1 STRIPES=1
  DEST=[134.79.198.146] TYPE=RETR CODE=226
```

- The usage log provides useful information such as:
  - Number of bytes transferred
  - Transfer start date
  - Transfer end date
  - Source/Destination
  - Configuration (streams/stripes settings)

Z. Liu, UVA

20

Session Analysis

- g = 1min, gap between transfers in a session
- current VC setup delay: 1 min

TABLE I: NCAR-NICS sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8,793 (bytes)	5,808.7	70,708.4	263,771.4	320,600	2,873,868.5

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	210.5	1,445	4,039	5,261	48,420

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.1 (bps)	298	468.3	506.1	682.2	4,227

TABLE II: SLAC-BNL sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
812 (bytes)	273	1,195	24,045	4,860	12,037,604

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	18.95	58.92	315.9	151.1	95,080

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003	22.57	112.8	130.4	183.1	2,560

- Largest session in SLAC-BNL: ~12TB @ 1.06Gbps
- Longest Session in NCAR-NICS: 2.4TB, 13 hrs and 24 mins, 410Mbps
- SLAC-BNL: 78% of transfers belong to sessions that would have lasted longer than 10 min

Throughput Variance

- Significant variance observed
  - NCAR-NICS: CV = 63.05%
  - SLAC-BNL: CV = 115.24%

TABLE I: NCAR-NICS sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8,793 (bytes)	5,808.7	70,708.4	263,771.4	320,600	2,873,868.5

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	210.5	1,445	4,039	5,261	48,420

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.1 (bps)	298	468.3	506.1	682.2	4,227

TABLE II: SLAC-BNL sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
812 (bytes)	273	1,195	24,045	4,860	12,037,604

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	18.95	58.92	315.9	151.1	95,080

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003	22.57	112.8	130.4	183.1	2,560

Throughput Variance

- Potential factors impacting throughput
  - GridFTP configuration
    - number of parallel TCP streams
    - number of stripes (servers)
    - reuse TCP connections ("-fast" option)
  - Competing network/server resources
    - link utilization
    - concurrent transfers (GridFTP and other apps)

Key findings

- Data set transfers create alpha flows (high-rate, large-sized)
  - 4 Gbps average for one flow on a 10 Gb/s link
  - Bursts can adversely impact r.t. flows
  - Even with high rates, durations are long enough for the 1-min VC setup delay overhead
- Throughput variance
  - Partially due to configuration differences
  - Mainly due to competition for CPU and disk resources rather than network resources

Engineering

- Our hypothesis is that we will need to design a co-scheduling system for server and network resources to reduce throughput variance
- Engineering: implement a prototype and test

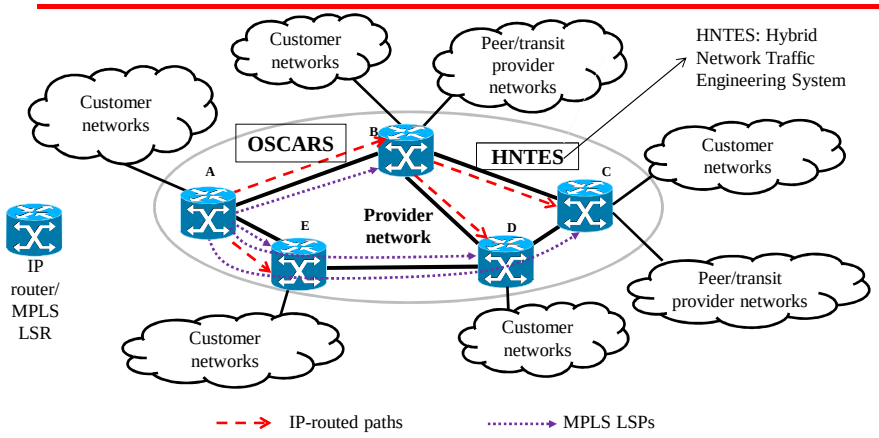
25

Our group's research

- Large dataset movement: GridFTP
- ➢ Hybrid network traffic engineering system
  - alpha-flow identification
  - redirection to virtual circuits
- Virtual circuit multicast transport protocol

26

Hybrid network traffic engineering system (HNTES)
- Intradomain identification/redirection of alpha flows



- OSCARS: On-demand Secure Circuits & Advanced Reservation System
- Hybrid: IP-routed + MPLS virtual circuits

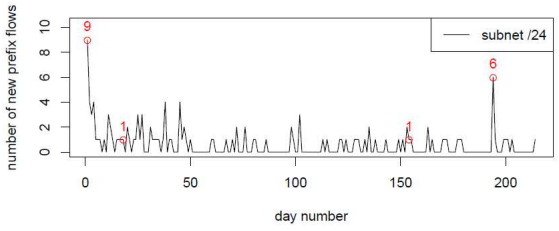
HNTES design

- alpha flow (high-rate, large sized) identification
  - scientists who move large sized datasets invest in high-end computers, high-speed disks, parallel file systems, and high access link speeds
  - data transfer nodes likely to have static IP addresses
  - NetFlow data over 7 months (May-Nov 2011) collected at ESnet router
  - find NetFlow reports where bytes sent in 1 minute > H bytes (1 GB)
  - save source/destination address prefixes (prefix flows)
  - age out entries if no flows appear for > 30 days
  - NetFlow data analysis to determine feasibility
- configure ingress routers to redirect alpha flows to traffic-engineered, QoS-controlled intra-domain paths
  - experimental studies of interaction between policing & scheduling schemes

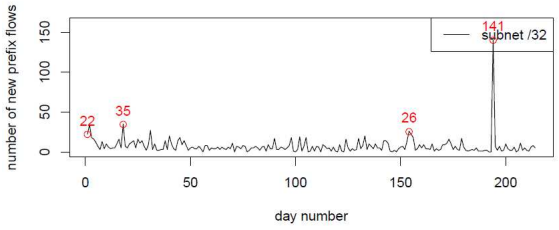
Z. Yan, C. Tracy, and M. Veeraraghavan, "A hybrid network traffic engineering system," in Proc. of IEEE 13th High Performance Switching and Routing (HPSR) 2012, June 24-27 2012.

Number of new prefix flows daily - Brookhaven (BNL) ESnet router

- For 193 out of 214 days only 0 or 1 new prefix flow
- When new collaborations start or new data transfer nodes are brought online, new prefix flows will occur

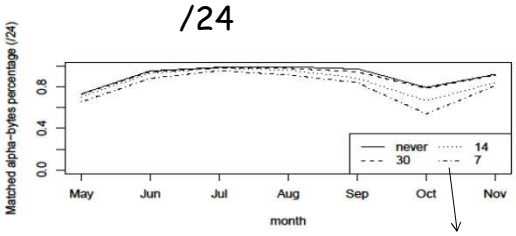


NetFlow: May-Nov. 2011



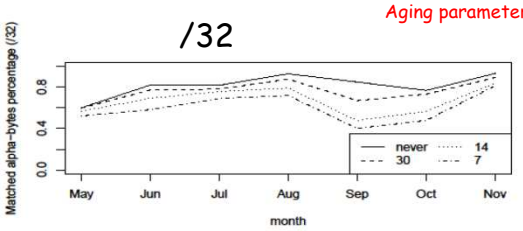
Percent of alpha bytes that would have been redirected

/24



Aging parameter

/32

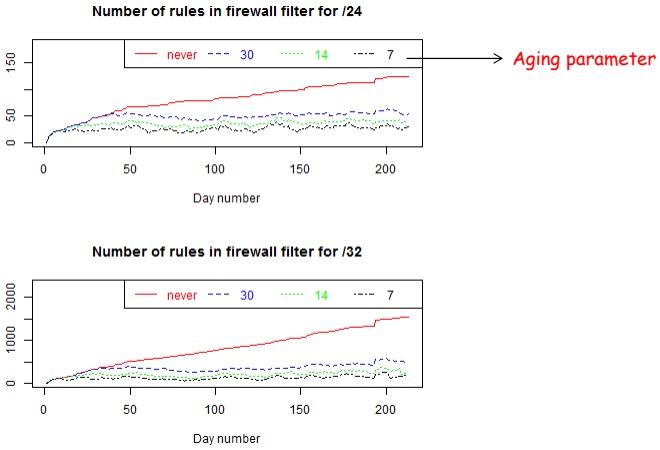


All 7 months:

Aging parameter	/24	/32
7	82%	67%
14	87%	73%
30	91%	82%
never	92%	86%

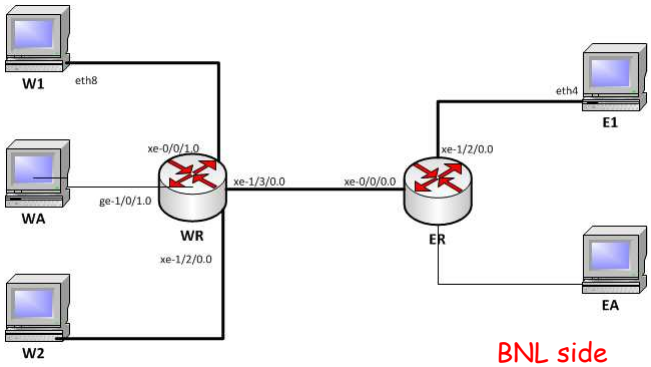
- When new collaborations start or new data transfer nodes are brought online, new prefix flows will occur, and so matched rates will drop

Size of firewall filter



Experimental studies to study router QoS mechanisms

- DOE-funded high-speed testbed
- High-performance computers



NEWY (6th Ave.)

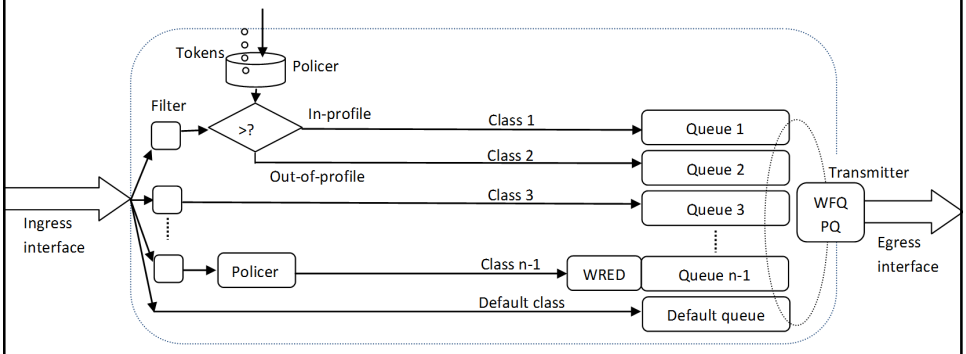
BNL side

Metro-area 10Gbps network

Router configurations

- Firewall filter:
  - alpha-flow based on /24 or /32 address prefixes (src and dst)
- Policing
  - classify out-of-profile packets as scavenger class and send to scavenger queue
  - Weighted Random Early Detection (WRED)
- Scheduling
  - Weighted fair queueing (WFQ)
  - Priority queueing (PQ)

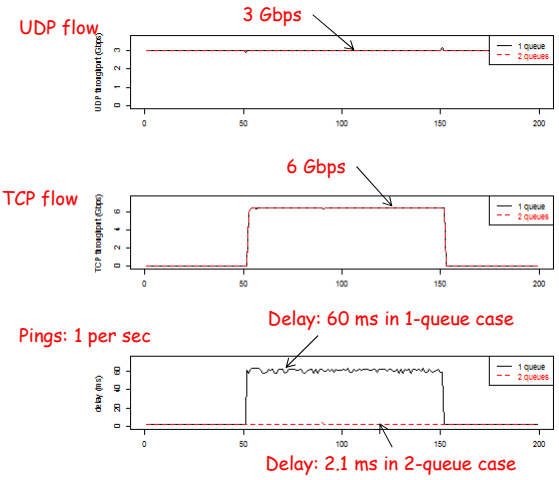
Policing on ingress scheduling on egress



Compare 3 configurations

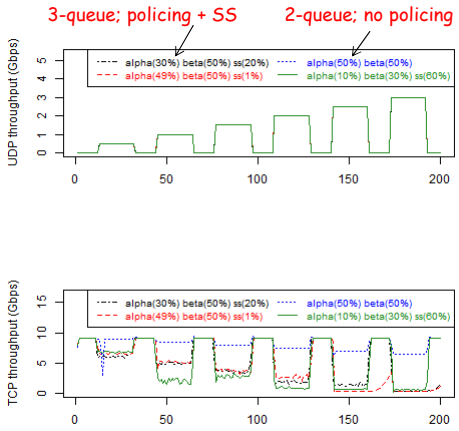
- 1-queue:
  - best-effort
  - all flows directed to same egress-side queue
- 2-queue: alpha and beta
  - scheduling-only (no policing)
    - WFQ +PQ
    - transmitter: shared in work conserving
    - buffer: strict partitioning
- 3-queue: alpha, beta, scavenger service (SS)
  - policing: > 1 Gbps sent to SS queue
  - scheduling; WFQ + PQ

Impact of alpha flows on real-time flows



- Impact on ping flow delay
  - significant in 1-queue configuration
  - negligible in 2-queue configuration
- Need separate virtual queue for alpha flow packets

Impact of policing: 3-queue case causes TCP throughput to drop



- When UDP rate is increased from 0.5 to 1 Gbps, significant drop in TCP throughput
- Why? out-of-sequence packets
- TCP fast-retransmit/fast recovery algorithm causes sending rate to drop by half
- Worst when alpha queue allocation is only 10%

Key findings

- Nodes capable of generating alpha flows have static public IP addresses
- Can leverage this fact in an offline HNTES design (nightly NetFlow analysis for alpha prefix ID determination)
- Configure ingress routers with firewall filters and WFQ/PQ 2-queue scheduling (alpha and beta); no policing
- Engr step: Integrate and test HNTES on ESnet
- Drawback of offline HNTES : new flows can cause trouble tickets!

Our group's research

- Large dataset movement: GridFTP
- Hybrid network traffic engineering system
- ➢ Virtual circuit multicast transport protocol

39

UCAR distributes climate data to 160 institutions

- Many different data feedtypes (30)
  - http://www.unidata.ucar.edu/software/ldm/ldm-current/basics/feedtypes/
- Of these:
  - CONDUIT: NCEP high-resolution model output
  - GEM: Canadian Meteorological Center GEM model output
  - NEXRAD2: NEXRAD Level-II radar data

40

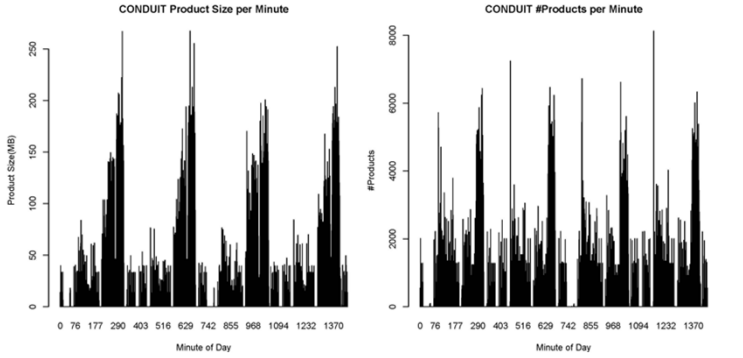
Real-time statistics

- http://www.unidata.ucar.edu/software/idd/rtstats/
- Ordered set of feedtypes (volume)
  - http://www.unidata.ucar.edu/cgi-bin/rtstats/rtstats_summary_volume?oliver.unidata.ucar.edu

41

CONDUIT data

- Installed and configured the LDM to receive CONDUIT data from UCAR: Jie Li
- Parsed and analyzed the log files for received data(9 sample days)
  - Peak throughput: 250 MB/minute (SD: 28.8 MB/minute)



42

Distribution structure

- Downloaded and parsed real-time statistics of the CONDUIT feed tree
- Data Distribution Topology of the CONDUIT feedtype
  - For the max fan-out of 104 receivers, the peak bandwidth requirement is 104 X 250 MB/minute ≈ 3.5Gbps
  - This is just for a single feedtype of a single application

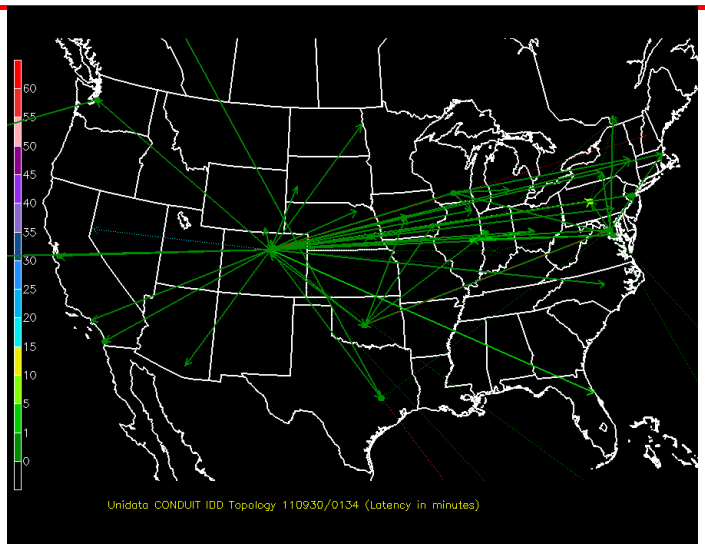
CONDUIT Feed Tree Topology Information

Parameter	Number
# Distinct Hosts	163
# Sender Hosts	57
# Receiver Hosts	141
Max. Fan-out Number	104*

* This maximum fan-out number comes from the UCAR site (idd.unidata.ucar.edu)

43

CONDUIT topology http://www.unidata.ucar.edu/cgi-bin/rtstats/rtstats_topogif?CONDUIT



44

Answer to question

- Different network service types
  - Static unicast VCs: unsuitable
    - Divide NCAR access link bandwidth between 104 subscribers: if 10 Gbps, then ~10 Mbps per subscriber
    - Subscribers would like to receive the data asap (low rate VC will increase latency)
  - Dynamic unicast VCs: unsuitable
    - For the worst-case fanout of 104, the total delay will be greater than with IP service, since for each receiver a new circuit needs to be set up, which can only be done after the transfer to the previous receiver is complete and the circuit to that receiver is released.
  - Multicast: can save bandwidth and computing resource

New options: multicast and P2P

- Multicast:
  - Hypothesis: total delay for distributing the data to the receivers will be lower for a given computing capacity of the upstream servers, or conversely, the same transfer delay can be achieved as with IP-routed service but with smaller upstream server computing capacity.
  - Negative: one or more slow receivers can slow down everyone
- P2P:
  - Hypothesis: transfer delay does not increase with number of receivers as with unicast TCP
  - Need to transmit data as soon as generated makes this less attractive

Multicast VCs

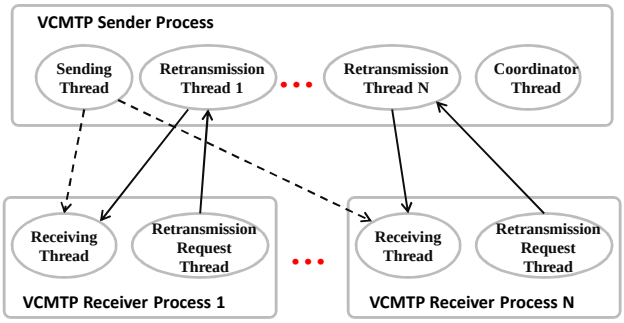
- Difference between multicast VC and multicast IP routed service
  - No congestion related losses on multicast VC
  - In-sequence delivery of packets
- Multicast Transport Protocols:
  - For IP-routed service: NORM and Raptor codes
  - No multicast transport protocol for VCs
    - To the best of our knowledge
  - Hence designing a reliable Virtual Circuit Multicast Transport Protocol (VCMTP)

VCMTP Key Design Concepts

- Message based transport protocol (in contrast to byte-stream based)
- Asynchronous API (RDMA-like): need time-out
- Automatic/application-driven receive mode
- Use Negative Acknowledgement (NACK) to avoid positive ACK-implosion problem
- Use reliable unicast connections for retransmissions
- Data retransmission occurs in parallel with data multicast to serve a continuous stream of messages
- Multicast groups with different send rates serve different groups of receivers

A Multithreaded Implementation

- VCMTP Sender Process creates a separate *Retransmission Thread* to communicate with the *Retransmission Requester Thread* on each receiver
- *Coordinator Thread* on VCMTPSender monitors the multicast status of each file and status of all receivers



49

U. Utah Emulab& UNM PRObE

- Testing VCMTP on NSF-funded clusters
- Shared by many projects
- Gives researchers root access
  - unlike supercomputing centers
- Automatically loads user image at time of reservation

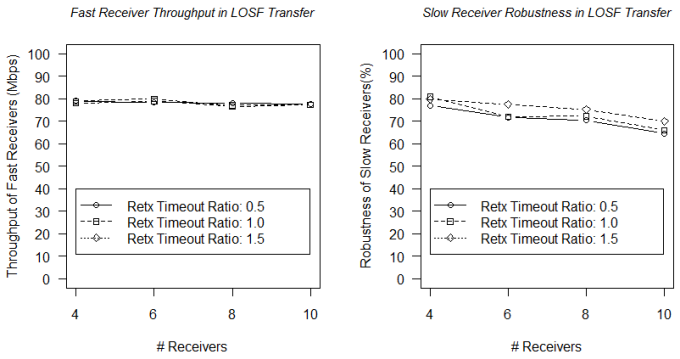
50

Multicast of continually generated files

- Scientific applications such as IDD need to distribute files as soon as they are created
- Goals
  - High throughput for fast receivers
  - High robustness for slow receivers (percentage of files received before sender timeout)
- Management-plane tracking:
  - slow receiver with low robustness switched to a multicast VC with a lower rate

Experiment Results

- The send rate was 100 Mbps with a traffic intensity of 0.9
- 50% of the receivers were slow receivers with a 5% loss rate
- Different retransmission timeout ratios were applied
- Each data point was the average value calculated from 30 repeated runs



Key findings

- VCMTP design:
  - For virtual circuits
  - Remote DMA (RDMA) like
  - To move continually created files, not single files
- Tradeoff via time-out value
  - Throughput for fast receivers
  - Robustness for slow receivers

Summary

- Brief flavor for the high-speed networking needs of the scientific community
- Approach:
  - Analyze data (collaborators: key: "science")
  - Prototype solutions to solve a problem ("engr")
  - Leverage community-wide shared resources (CRI and MRI NSF funding)
- Slides/comments? mvee@virginia.edu